

ALEXSANDRO PESSOA

Senior AI Platform Engineer · Agentic Systems · Data Platforms · Developer Enablement

a14xdev.github.io

PROFILE

AI & Platform Engineer who turns ambiguous GenAI architectures into deployable, observable, reusable systems: typed LLM workflows, agent control planes, vLLM runtimes, and developer tooling. Measured impact on cost, performance, and delivery speed.

EXPERIENCE

OCT 2025 — PRESENT
Senior AI Platform Engineer Accenture

Enterprise data & AI platform

I work with terabyte-scale, business-critical Billing data for two of Brazil's largest telecommunications operators. I conduct technical investigations and structure analyses and recommendations used by managers in the project's decision-making process, connecting observed failures to architectural and implementation direction. I lead architecture, developer enablement, and technical direction for a confidential enterprise data/AI platform built on Azure and Databricks. I modernize legacy flows with deterministic pipelines and build auditable GenAI foundations through explicit contracts, traceable validation, and isolated extensions that let developers add rules safely and autonomously. The selected research below makes public the same practice I apply in confidential work: measuring unsolved failure modes and translating them into explicit state, deterministic boundaries, and traceable validation for probabilistic multi-agent systems. Client identities and internal metrics are withheld.

Engineering mentorship program

Designed and ran a technical mentorship program to raise the engineering quality bar across squads: modern Python, explicit typing, design patterns, and applied ML theory for GenAI. Tracked mentee progress with commit and delivery metrics.

OCT 2024 — 2025
AI Platform Engineer Accenture

Agentic document-transformation platform

Drove substantial implementation and evolution of an agentic document-transformation platform, including a reusable typed multi-stage LLM framework with retries, timeouts, partial failures, and telemetry. A state-machine orchestration framework (FastAPI/Pydantic) sped developer onboarding ~60%; lazy-loading cut LLM token spend ~30% and runtime memory ~80%; YAML-driven CI/CD reduced deployment time ~80%.

GenAI data pipelines & platform

Built high-throughput ETL pipelines (Polars/Python) that improved throughput ~50%, and cut BigQuery cost ~90% with a metadata-driven ORM. Delivered GenAI validation tooling and RAG services on containerized GCP infrastructure (Cloud Functions, Pub/Sub).

vLLM-based inference platform

Implemented runtime, deployment, observability, accounting integration, and operational automation for a vLLM-based inference platform; initial design constraints were shared.

INDEPENDENT ENGINEERING PROJECTS

Crime Alley CV Deterministic coding-agent control plane with typed transactional state, locks, recoverable commits, append-only event journals, offline replay, MCP debugging, and Docker/Podman isolation.

Alex Tavern ecosystem Source-visible proprietary multi-agent system with explicit state and information boundaries, compaction, replay/debug tooling, and a curated plugin hub, including a RAG plugin.

TECHNICAL FOCUS

PYTHON

LLM ENGINEERING

AGENTIC AI

RAG

LLMOPS

VLLM INFERENCE

DATA ENGINEERING

GCP

AZURE

DATABRICKS

TYPED ASYNC LLM WORKFLOWS

DETERMINISTIC CONTROL PLANES

OBSERVABILITY

DEVELOPER TOOLING

ADVANCED PYTHON SYSTEMS

LANGUAGES

Portuguese	NATIVE
English	C2
goFLUENT ECEFR SET	Apr 2025

MENTORED BY

MS

Marcelo Sabadini
Senior AI Solutions Consultant — GenAI architecture, LLM strategy, and executive-level AI transformation.

TR

Tiago Pereira Ramos
Senior Solutions Architect — agentic systems, hyperautomation, and enterprise AI governance.

To you—person or AI—reading this dedication: science and engineering advance through the knowledge of those who came before us and those who share it today. I therefore recognize the people who contributed, directly or indirectly, to my knowledge and professional development. This continuity does not diminish my independence: I author and take full responsibility for every decision, work, and result presented here.

EDUCATION & CERTIFICATIONS

B.Eng. Materials Engineering, UFCA (2023)

Materials Engineering studies, UFSC (2016–2019)

Machine Learning Specialization, Stanford Online & DeepLearning.AI (2024)

People Leadership Credential L1, Accenture (2025)

Consumer Electronics Industry Foundations, Primerli (2025)

PromptNest Typed async Python nested map/reduce orchestration for OpenAI, Azure OpenAI, LangChain, LangGraph, CrewAI, and custom LLM runtimes, with bounded backpressure, provider rate/token limits, retries, partial recovery, and durable checkpoints; simulated-provider benchmark: 10,000 jobs, 0.022 ms/job p95 orchestration overhead, 2.46 ms batch p99, and 1.46 ms TTFT p99 (adapter stream start to first non-empty text delta) across 100 streams.

SceneQueue Local-first ComfyUI storyboard with persistent sequence editing, hybrid job synchronization, and PWA/mobile UX.

ReImagineX Local-first image-lineage product on ComfyUI with atomic persistence, resilient job tracking, and mobile/PWA UX.

SELECTED ENGINEERING RESEARCH

Alex Tavern I — Structural Epistemic Consistency in Long-Running Multi-Agent Roleplay

Investigated observer-relative knowledge as a system invariant. In controlled replay of a real-session payload, perspective-safe context projection reduced identity leakage from 7/13 to 0/13 runs.

Alex Tavern II — Structural Narrative Time in Long-Running Multi-Agent Roleplay

Diagnosed semantic stagnation through controlled payload replay: abstract anti-loop instructions failed to break the cycle in 3/3 runs, concrete world events broke it in 2/3, and a concrete Director beat in 3/3. Proposed an authoritative tick-based world clock; validation remains in progress.